Data Science Portofolio

SQL In BigQuery

Yodi Ramadhani ALfariz

yodialfa.github.io

Subject

- How to make project and add dataset in BigQuery
- How to Connect BigQuery using Collaboratory
- How to access BigQuery using SQL in Collaboratory

You have to register first in cloud.google.com

for detail material and notebook, you can visit my github yodialfa/bigQuery-BeeCycle: SQL in BigQuery (github.com)

What is BigQuery ?

BigQuery is a fully managed enterprise data warehouse that helps you manage and analyze your data with built-in features like machine learning, geospatial analysis, and business intelligence. BigQuery's serverless architecture lets you use SQL queries to answer your organization's biggest questions with zero infrastructure management. BigQuery's scalable, distributed analysis engine lets you query terabytes in seconds and petabytes in minutes.

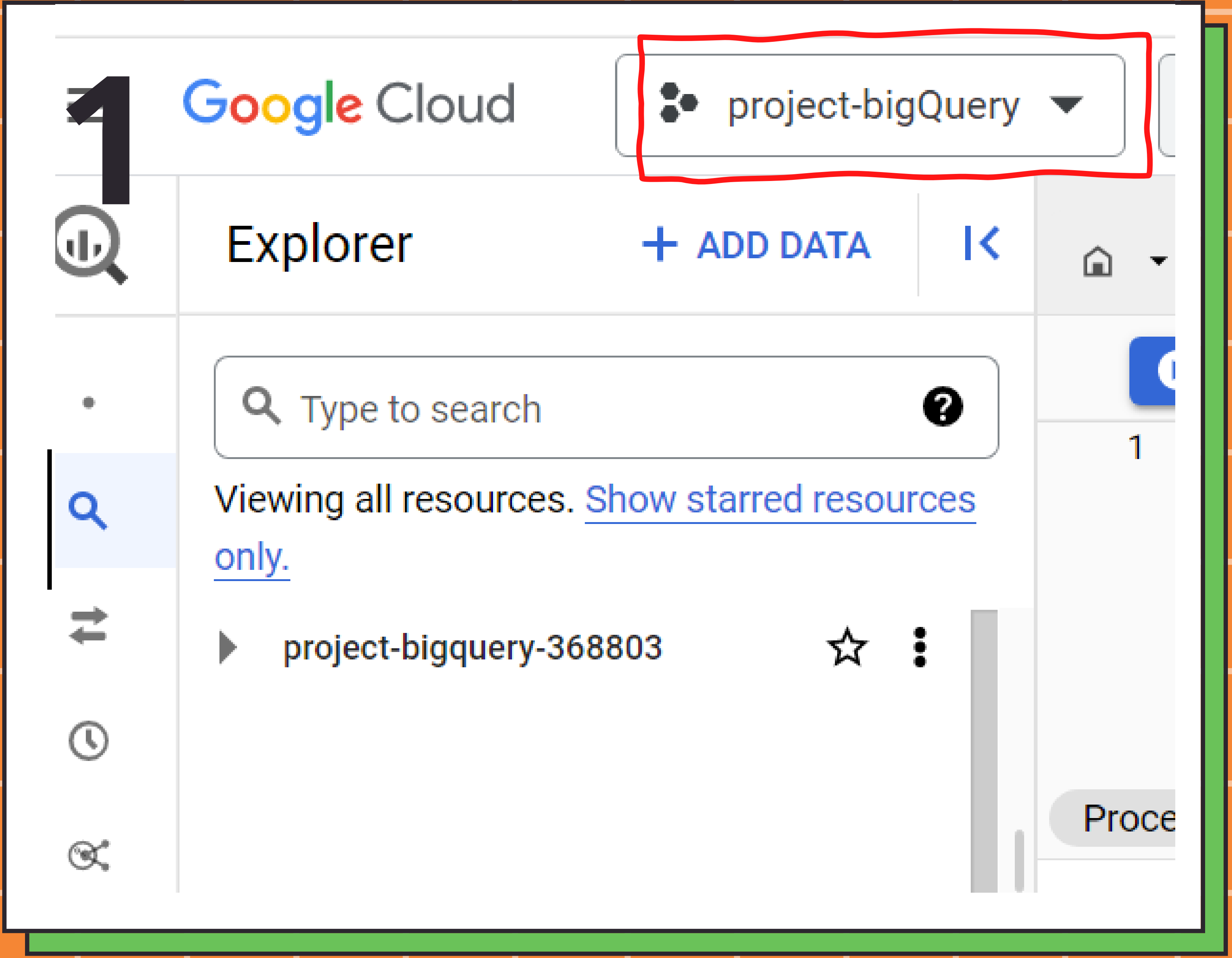
How to Make Project in BigQuery ?

1. Make Sure that you have Google Account
2. Visit website https://console.cloud.google.com/ and you have to register first to access bigquery
3. Klik on my-project beside Google Cloud logo
4. Klik new project

Make BigQuery Project

Make BigQuery Project

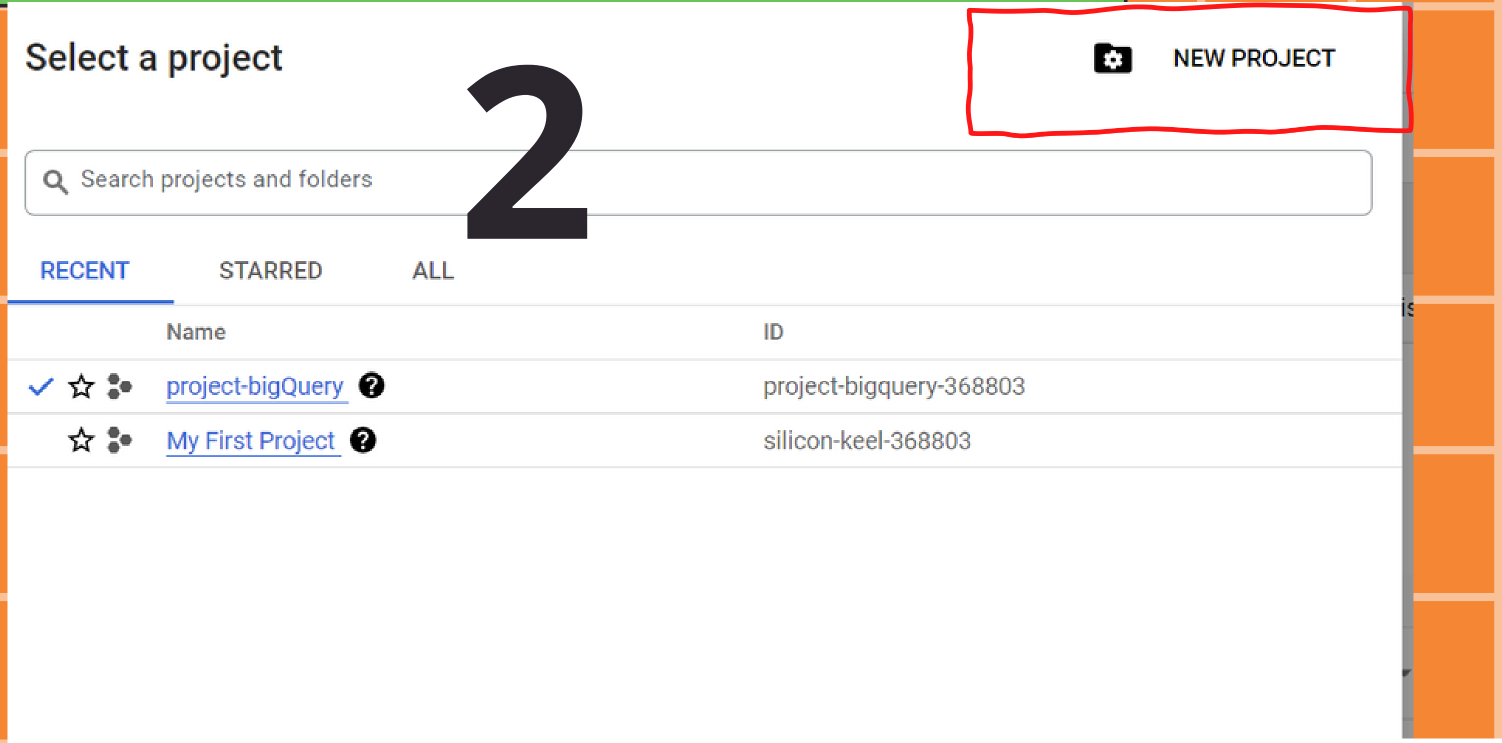
1



Viewing all resources. Show starred resources only.

project-bigquery-368803

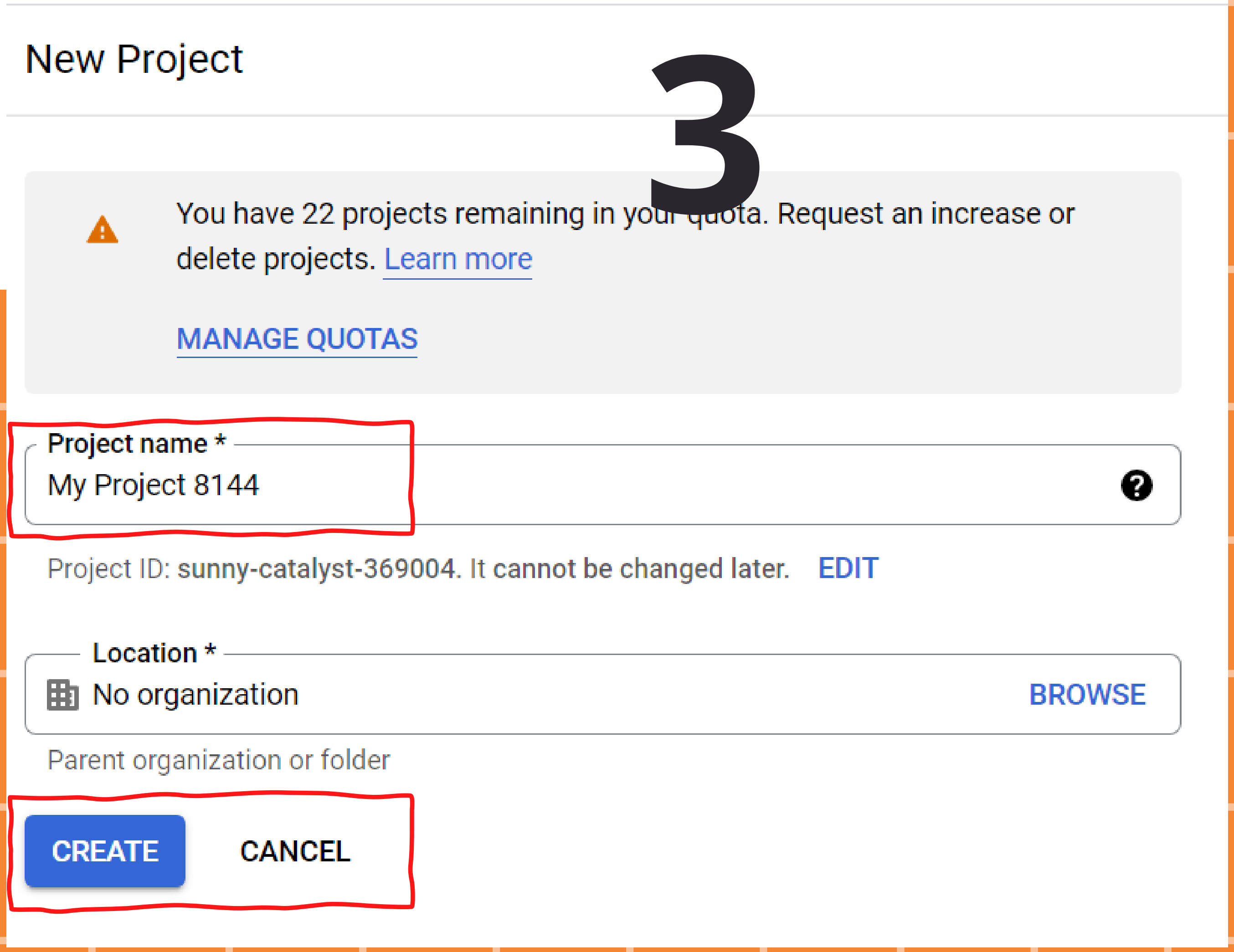
2



Select a project

	Name	ID
✓ ☆	project-bigQuery	project-bigquery-368803
☆	My First Project	silicon-keel-368803

3



New Project

You have 22 projects remaining in your quota. Request an increase or delete projects. Learn more

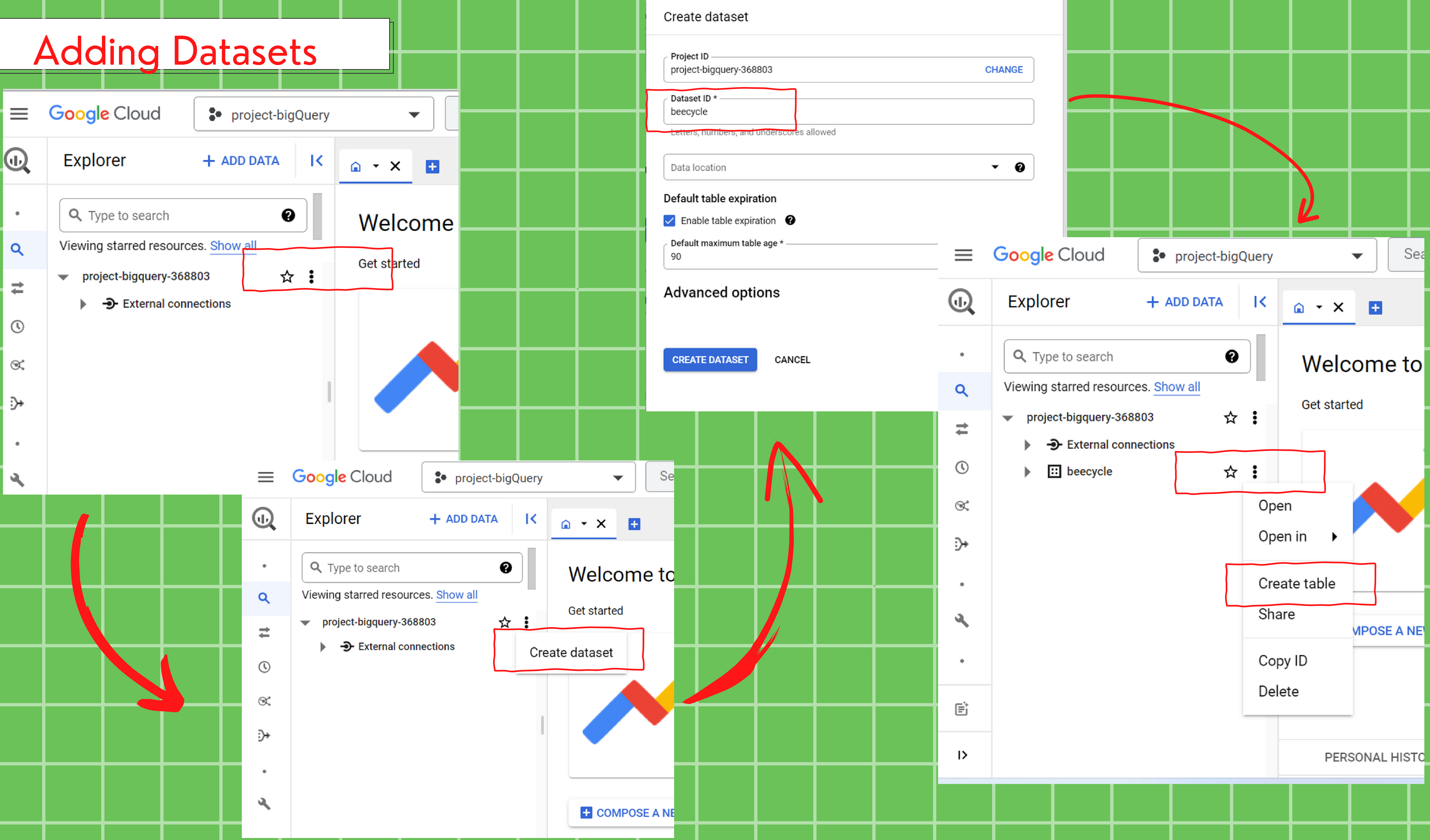
Project name *: My Project 8144

Project ID: sunny-catalyst-369004. It cannot be changed later. EDIT

Location *: No organization

Parent organization or folder

Adding Datasets



Google Cloud project-bigQuery

Explorer + ADD DATA

Type to search

Viewing starred resources. Show all

project-bigquery-368803

External connections

Welcome

Get started

Google Cloud project-bigQuery

Explorer + ADD DATA

Type to search

Viewing starred resources. Show all

project-bigquery-368803

External connections

Create dataset

Welcome to

Get started

COMPOSE A NE

Create dataset

Project ID: project-bigquery-368803 CHANGE

Dataset ID *: beecycle

Letters, numbers, and underscores allowed

Data location

Default table expiration

☑ Enable table expiration

Default maximum table age *: 90

Advanced options

CREATE DATASET CANCEL

Google Cloud project-bigQuery

Explorer + ADD DATA

Type to search

Viewing starred resources. Show all

project-bigquery-368803

External connections

beecycle

Open

Open in

Create table

Share

Copy ID

Delete

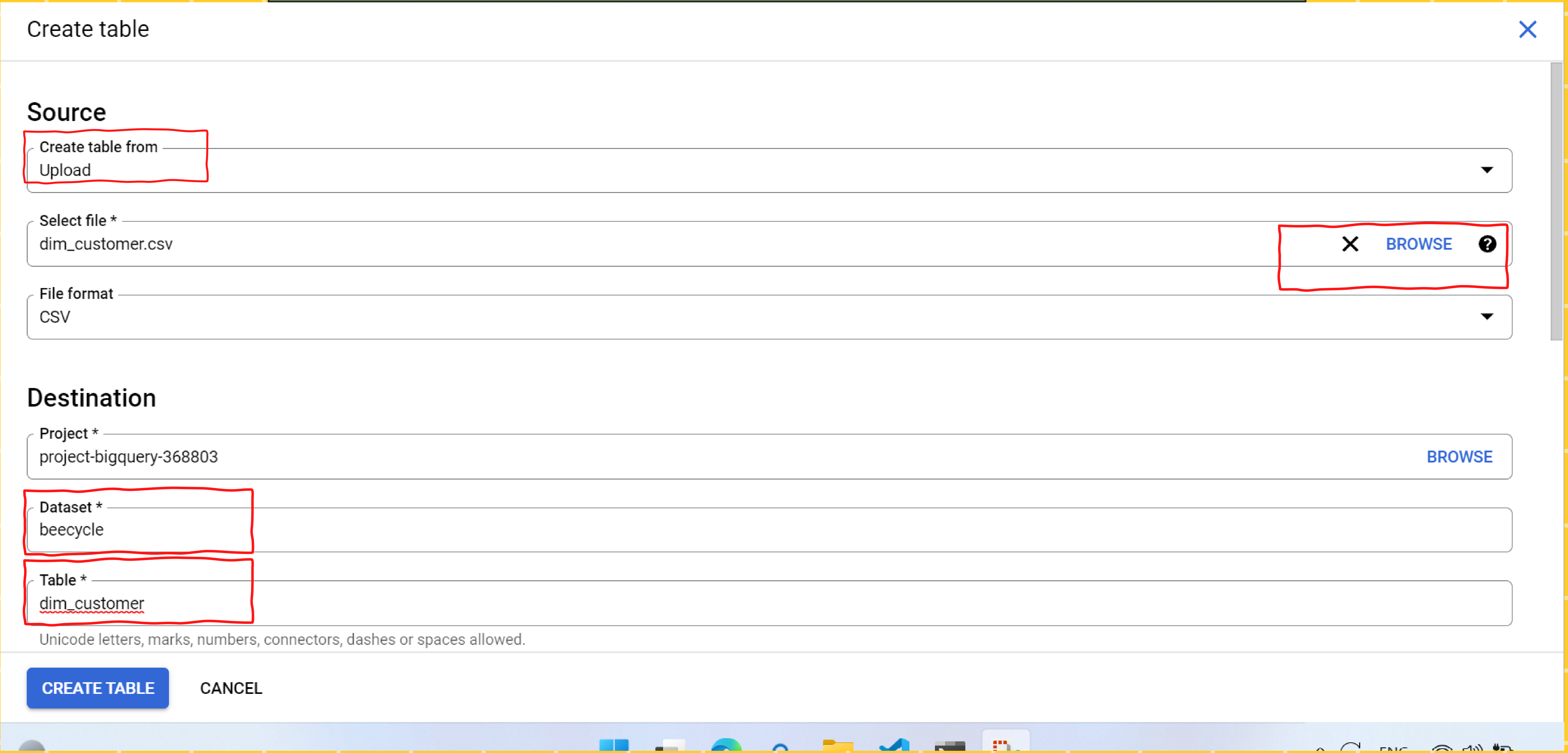
Welcome to

Get started

MPOSE A NE

PERSONAL HISTO

Adding Table Using Source



Create table

Source

Create table from
Upload

Select file *
dim_customer.csv

File format
CSV

Destination

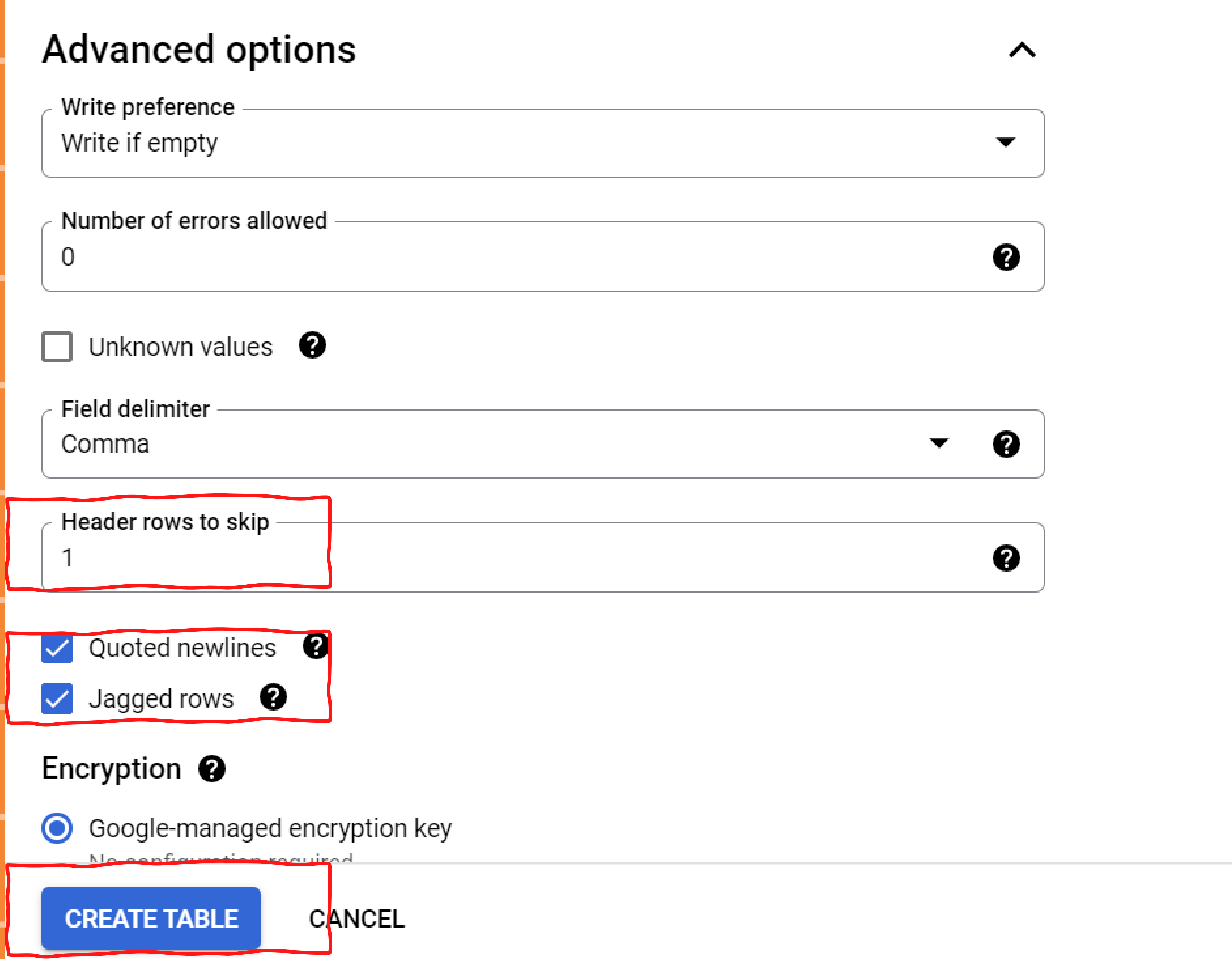
Project *
project-bigquery-368803

Dataset *
beecycle

Table *
dim_customer

Unicode letters, marks, numbers, connectors, dashes or spaces allowed.

Adding Datasets



Advanced options

Write preference: Write if empty

Number of errors allowed: 0

□ Unknown values

Field delimiter: Comma

Header rows to skip: 1

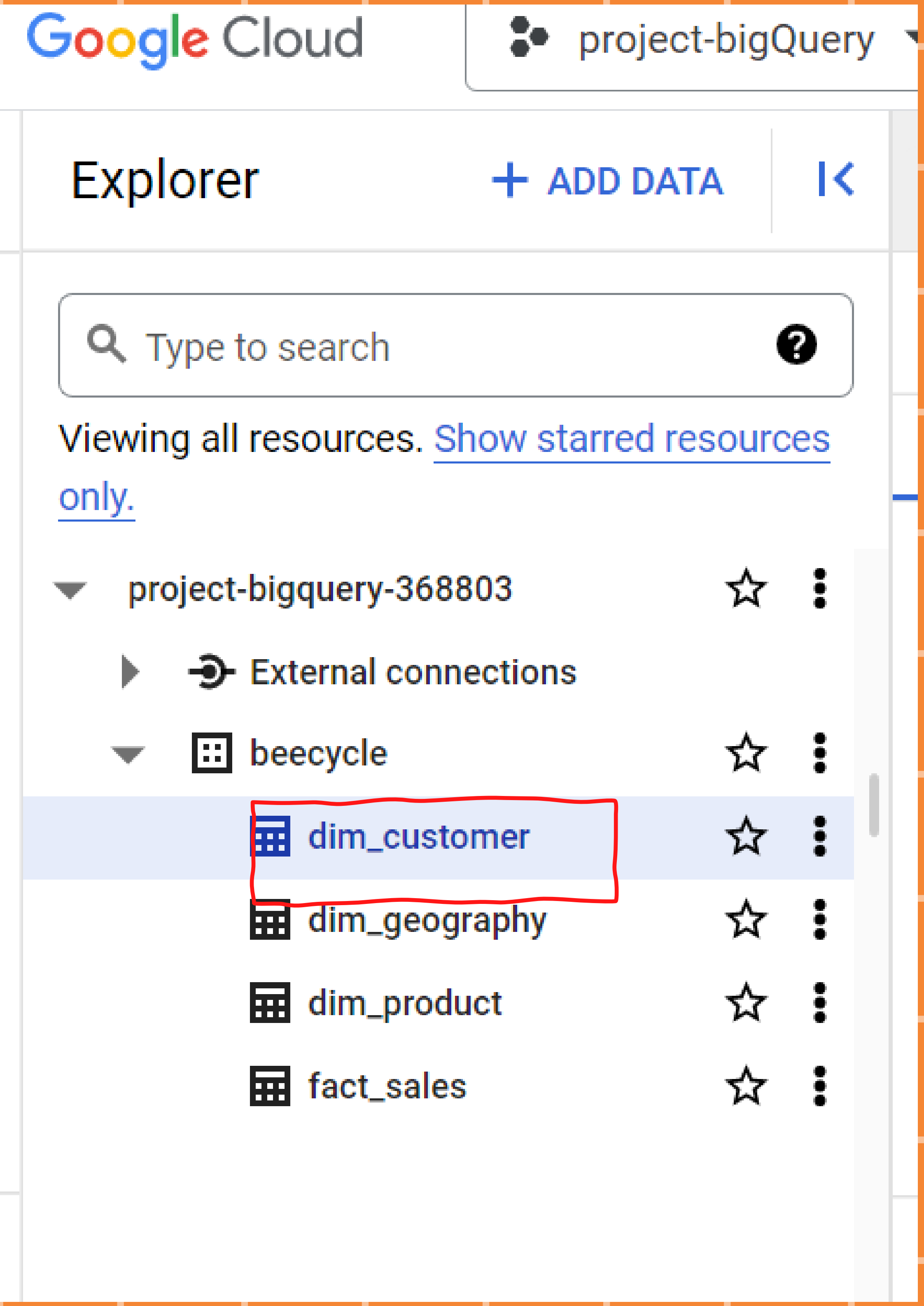
☑ Quoted newlines

☑ Jagged rows

Encryption

● Google-managed encryption key

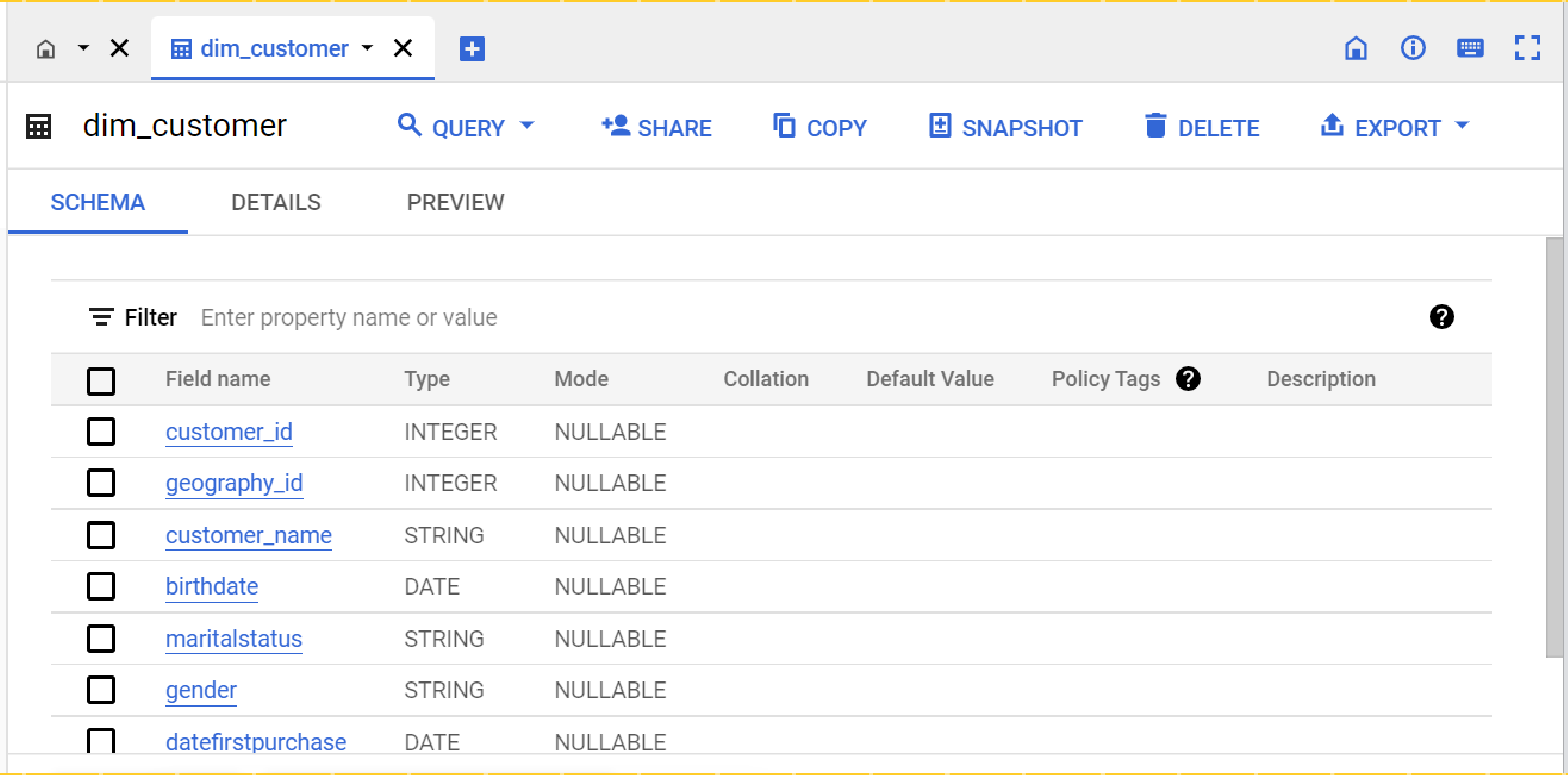
No configuration required



Viewing all resources. Show starred resources only.

- project-bigquery-368803
  - External connections
  - beecycle
    - dim_customer
    - dim_geography
    - dim_product
    - fact_sales

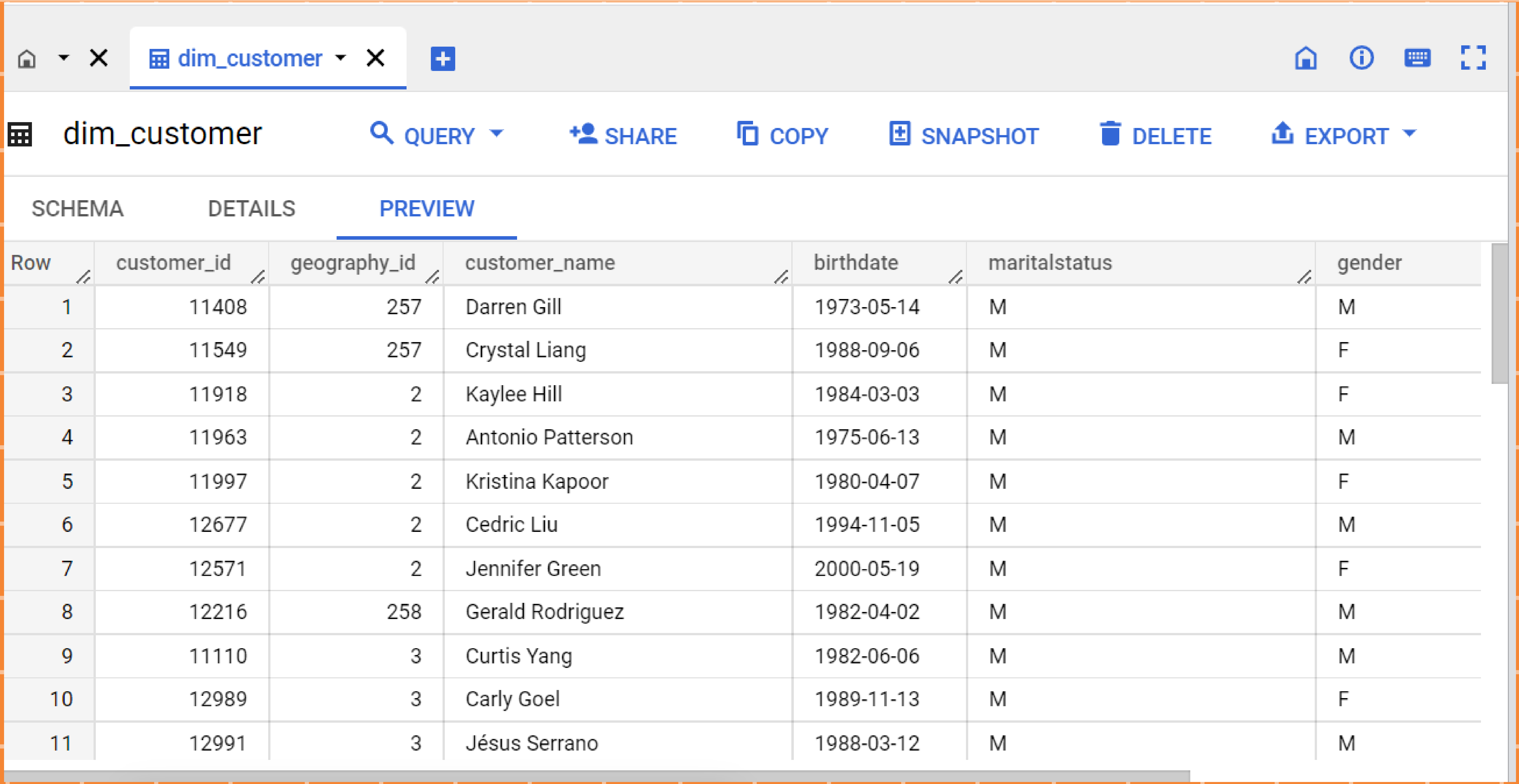
Table Schema Preview



dim_customer

□	Field name	Type	Mode	Collation	Default Value	Policy Tags	Description
□	customer_id	INTEGER	NULLABLE				
□	geography_id	INTEGER	NULLABLE				
□	customer_name	STRING	NULLABLE				
□	birthdate	DATE	NULLABLE				
□	maritalstatus	STRING	NULLABLE				
□	gender	STRING	NULLABLE				
□	datefirstpurchase	DATE	NULLABLE				

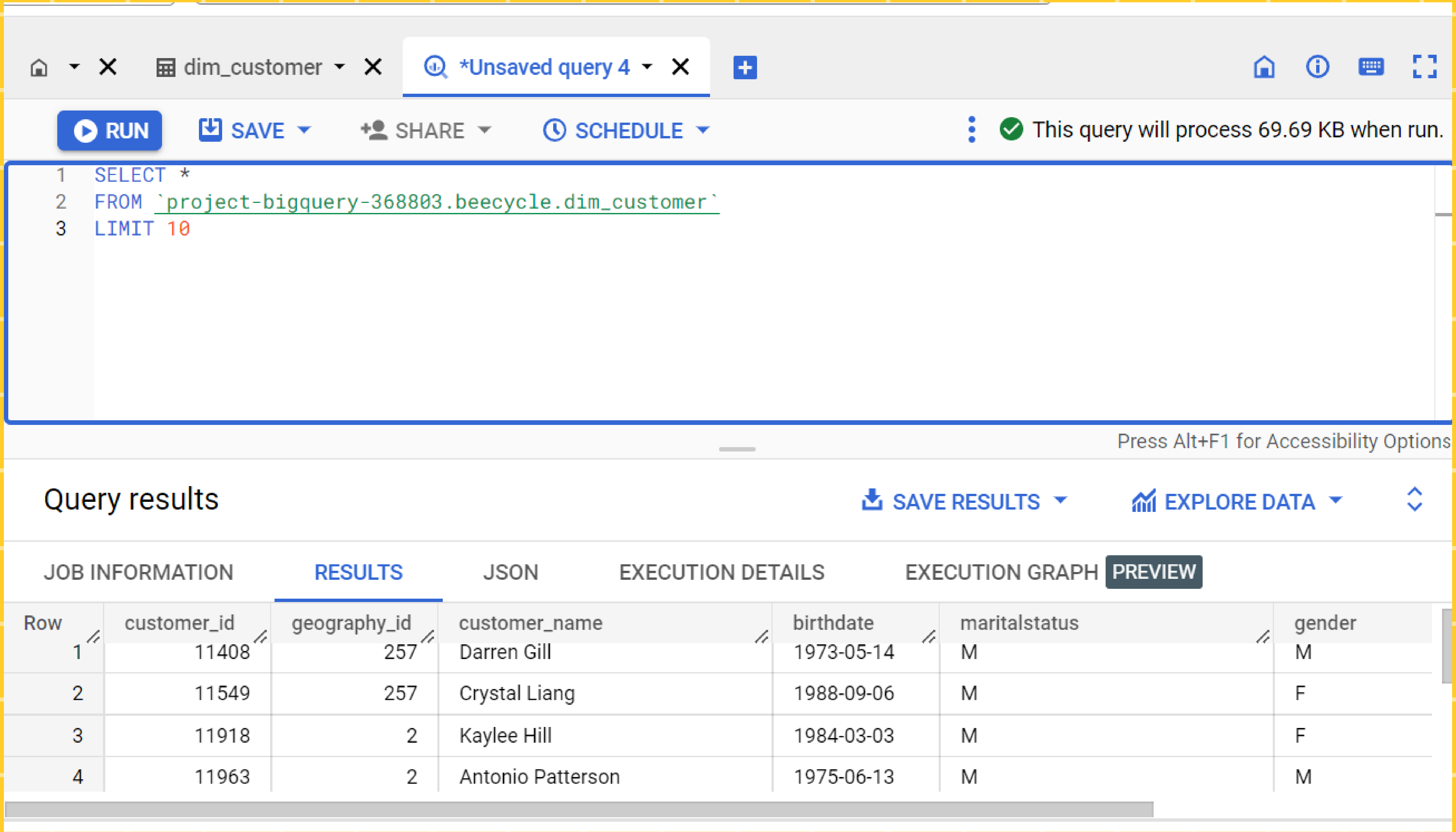
Data Preview



dim_customer

Row	customer_id	geography_id	customer_name	birthdate	maritalstatus	gender
1	11408	257	Darren Gill	1973-05-14	M	M
2	11549	257	Crystal Liang	1988-09-06	M	F
3	11918	2	Kaylee Hill	1984-03-03	M	F
4	11963	2	Antonio Patterson	1975-06-13	M	M
5	11997	2	Kristina Kapoor	1980-04-07	M	F
6	12677	2	Cedric Liu	1994-11-05	M	M
7	12571	2	Jennifer Green	2000-05-19	M	F
8	12216	258	Gerald Rodriguez	1982-04-02	M	M
9	11110	3	Curtis Yang	1982-06-06	M	M
10	12989	3	Carly Goel	1989-11-13	M	F
11	12991	3	Jésus Serrano	1988-03-12	M	M

Try SQL In BigQueryConsole



This query will process 69.69 KB when run.

```sql
SELECT *
FROM `project-bigquery-368803.beecycle.dim_customer`
LIMIT 10
```

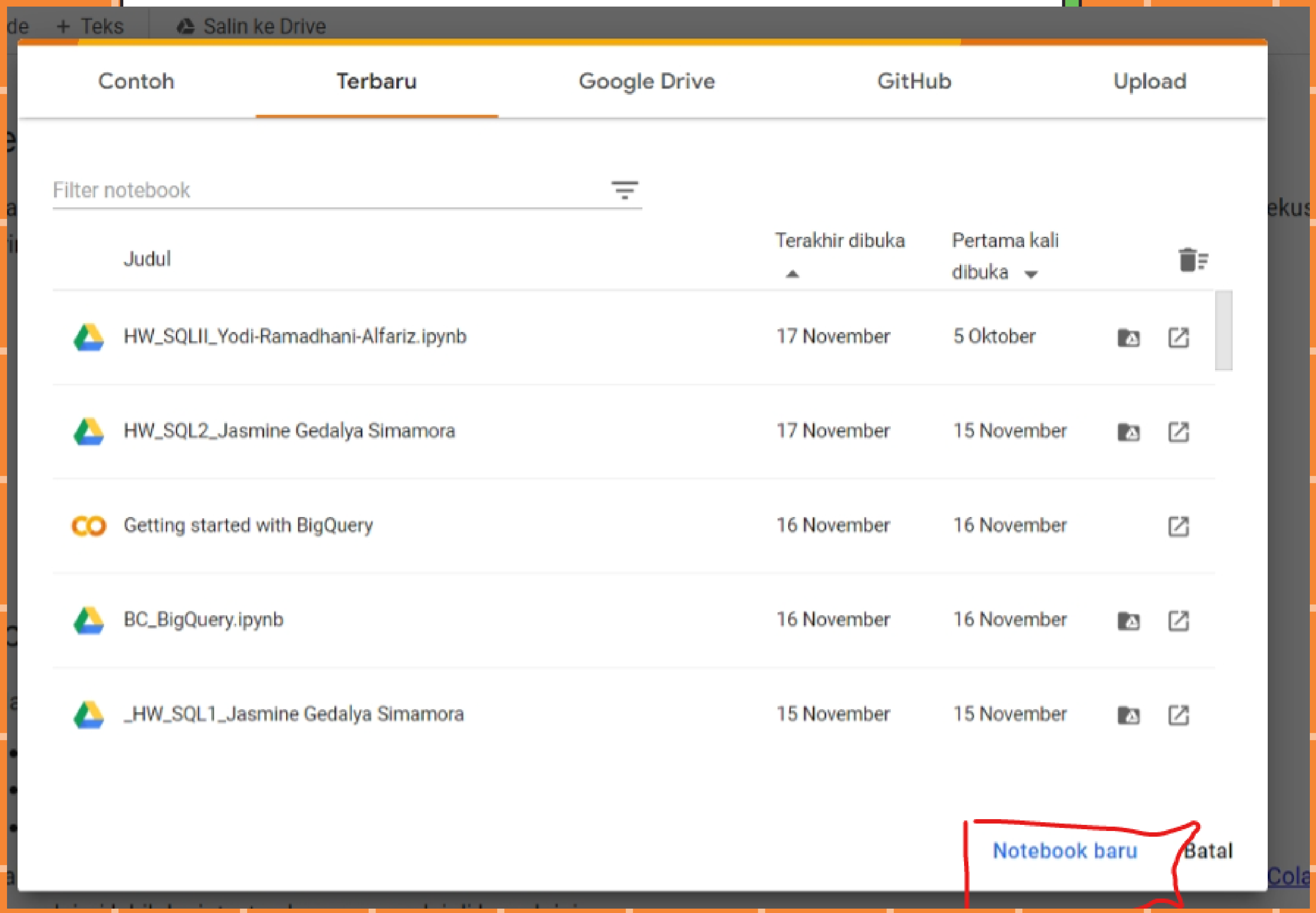
Query results

Row	customer_id	geography_id	customer_name	birthdate	maritalstatus	gender
1	11408	257	Darren Gill	1973-05-14	M	M
2	11549	257	Crystal Liang	1988-09-06	M	F
3	11918	2	Kaylee Hill	1984-03-03	M	F
4	11963	2	Antonio Patterson	1975-06-13	M	M

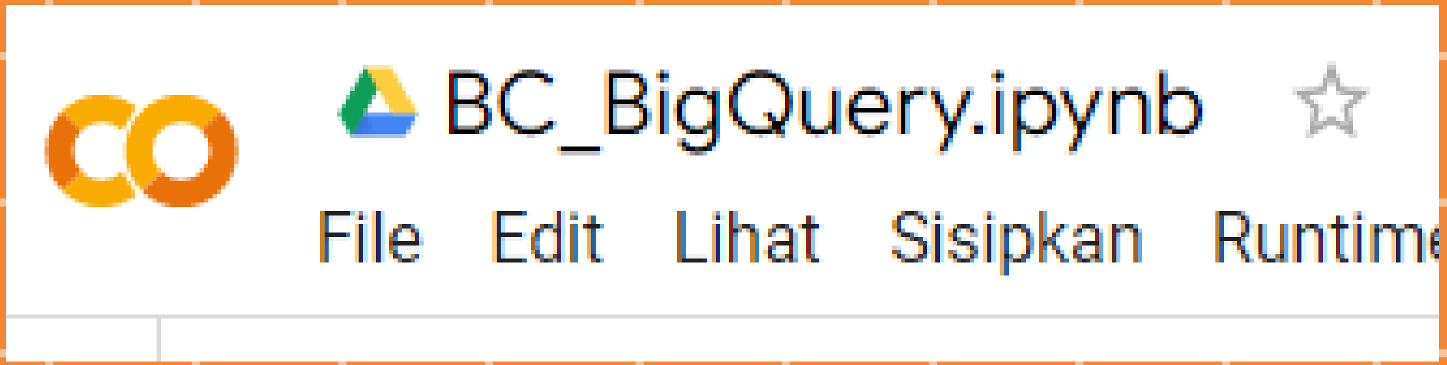
Connect BigQuery Using Google Colab

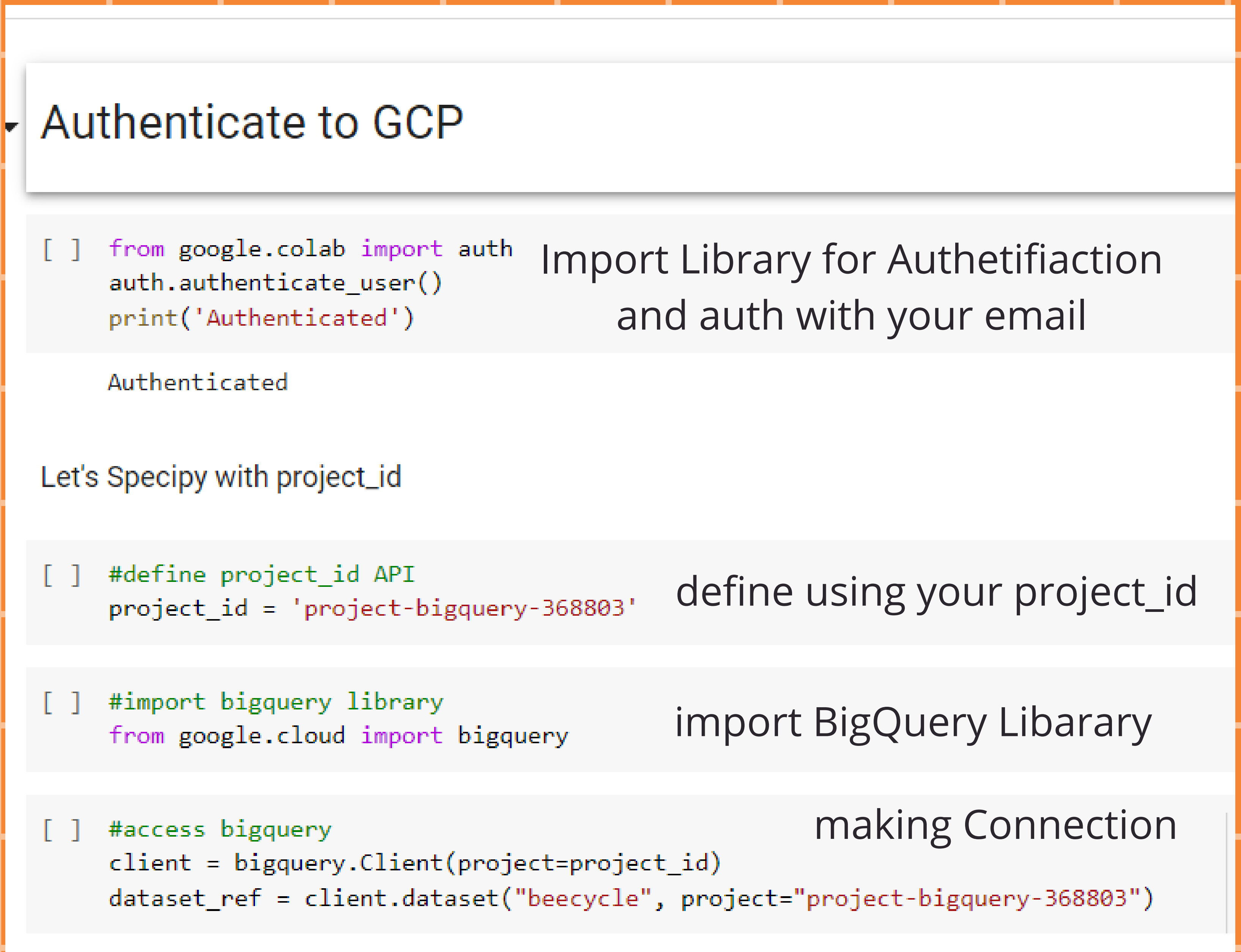
Make BigQuery Project

Access colab.google.com first



Judul	Terakhir dibuka	Pertama kali dibuka
HW_SQLII_Yodi-Ramadhani-Alfariz.ipynb	17 November	5 Oktober
HW_SQL2_Jasmine Gedalya Simamora	17 November	15 November
Getting started with BigQuery	16 November	16 November
BC_BigQuery.ipynb	16 November	16 November
_HW_SQL1_Jasmine Gedalya Simamora	15 November	15 November





Authenticate to GCP

```
from google.colab import auth
auth.authenticate_user()
print('Authenticated')
```

Import Library for Authetifiaction and auth with your email

```
Authenticated
```

Let's Specipy with project_id

```
#define project_id API
project_id = 'project-bigquery-368803'
```

define using your project_id

```
#import bigquery library
from google.cloud import bigquery
```

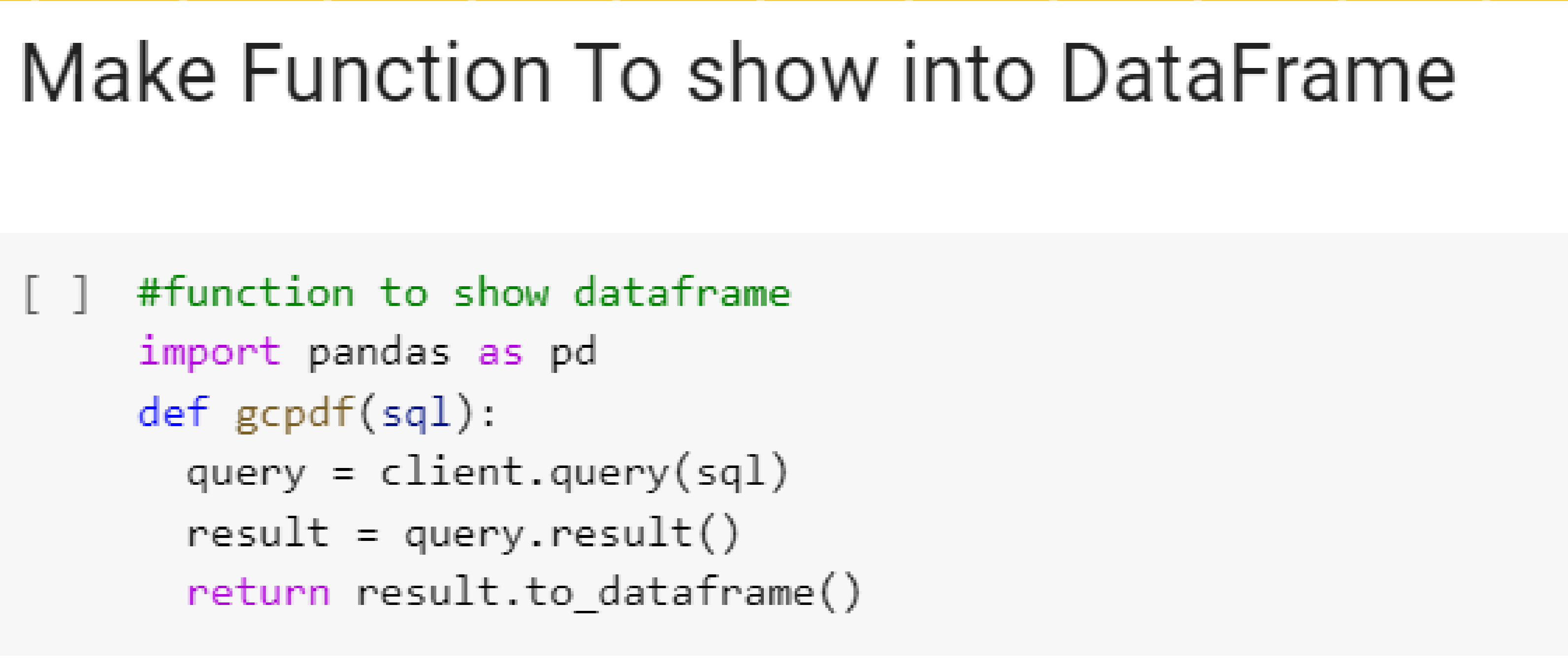
import BigQuery Libarary

```
#access bigquery
client = bigquery.Client(project=project_id)
dataset_ref = client.dataset("beecycle", project="project-bigquery-368803")
```

making Connection

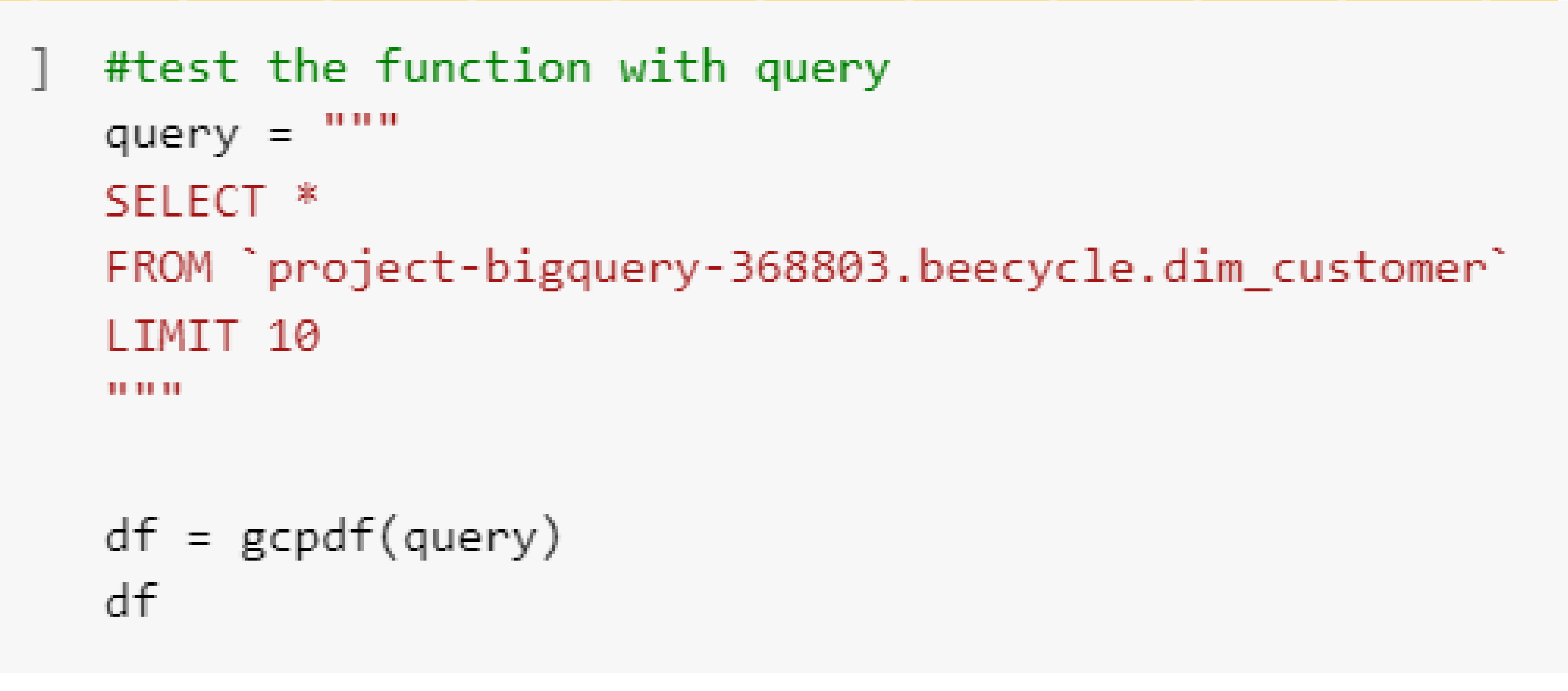
Access SQL in Colab

Making Function to Show the Data

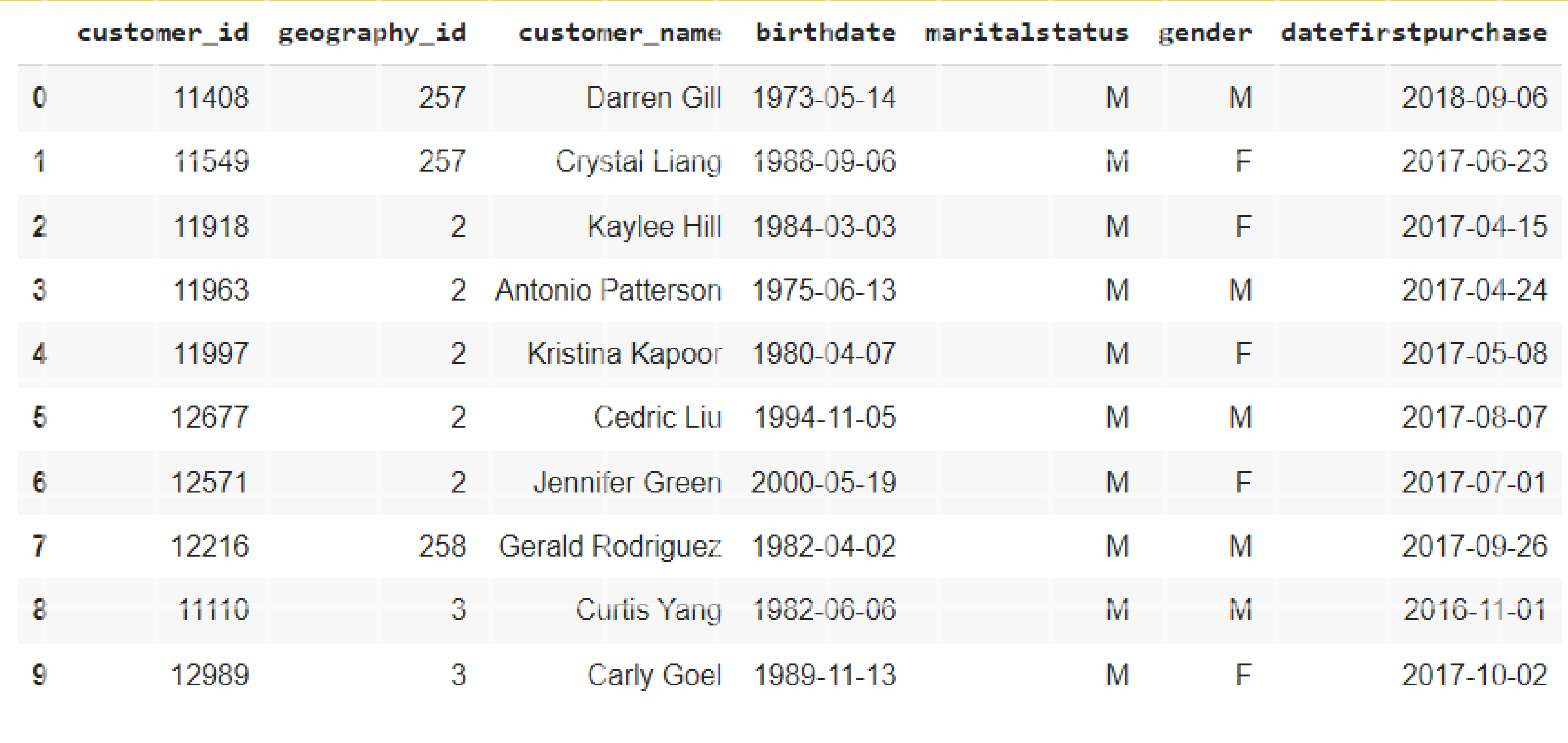


Make Function To show into DataFrame

```python
#function to show dataframe
import pandas as pd
def gcpdf(sql):
  query = client.query(sql)
  result = query.result()
  return result.to_dataframe()
```

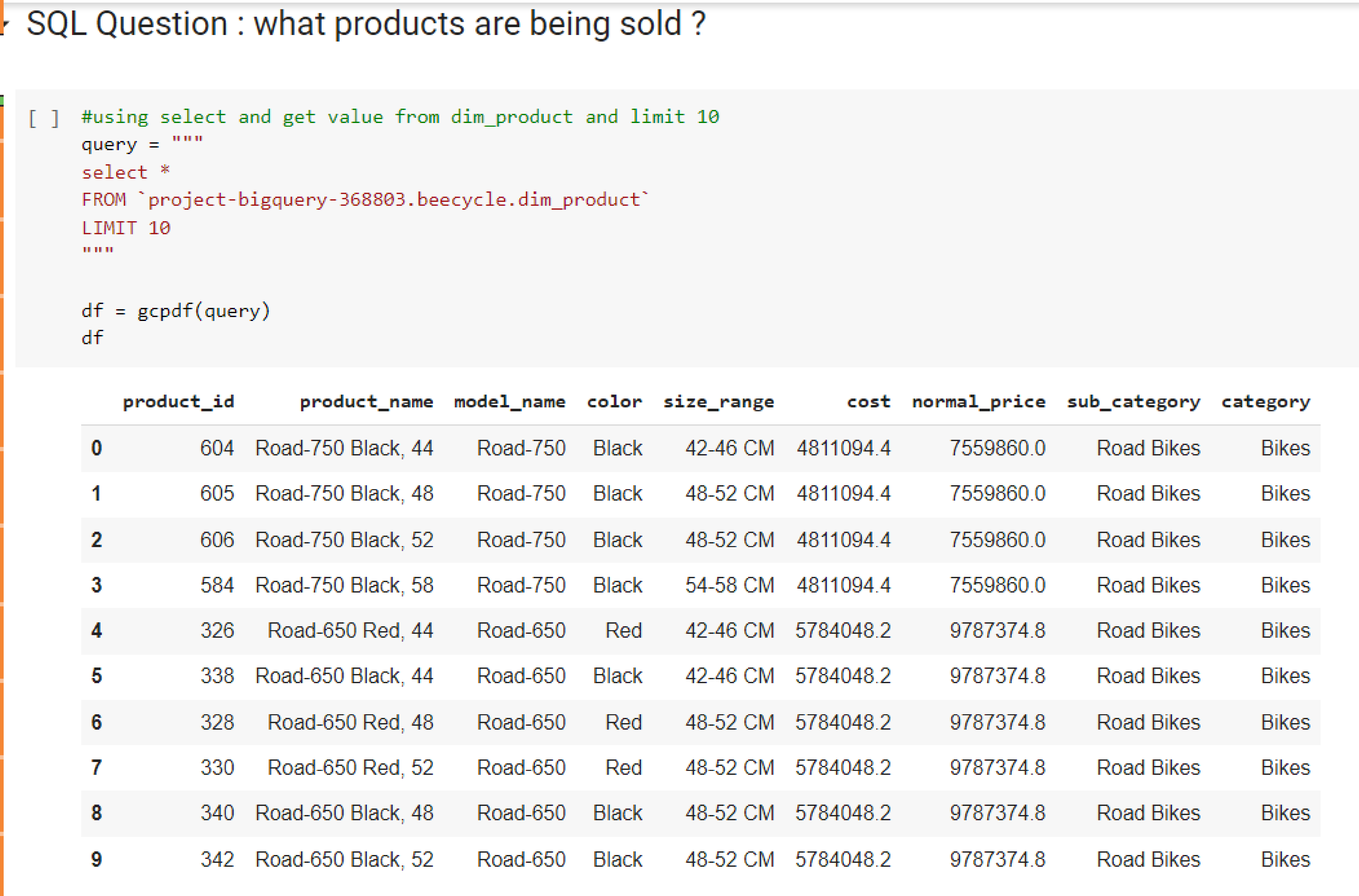


```python
#test the function with query
query = """
SELECT *
FROM `project-bigquery-368803.beecycle.dim_customer`
LIMIT 10
"""

df = gcpdf(query)
df
```



	customer_id	geography_id	customer_name	birthdate	maritalstatus	gender	datefirstpurchase
0	11408	257	Darren Gill	1973-05-14	M	M	2018-09-06
1	11549	257	Crystal Liang	1988-09-06	M	F	2017-06-23
2	11918	2	Kaylee Hill	1984-03-03	M	F	2017-04-15
3	11963	2	Antonio Patterson	1975-06-13	M	M	2017-04-24
4	11997	2	Kristina Kapoor	1980-04-07	M	F	2017-05-08
5	12677	2	Cedric Liu	1994-11-05	M	M	2017-08-07
6	12571	2	Jennifer Green	2000-05-19	M	F	2017-07-01
7	12216	258	Gerald Rodriguez	1982-04-02	M	M	2017-09-26
8	11110	3	Curtis Yang	1982-06-06	M	M	2016-11-01
9	12989	3	Carly Goel	1989-11-13	M	F	2017-10-02

Try SQL in Colab Using Question

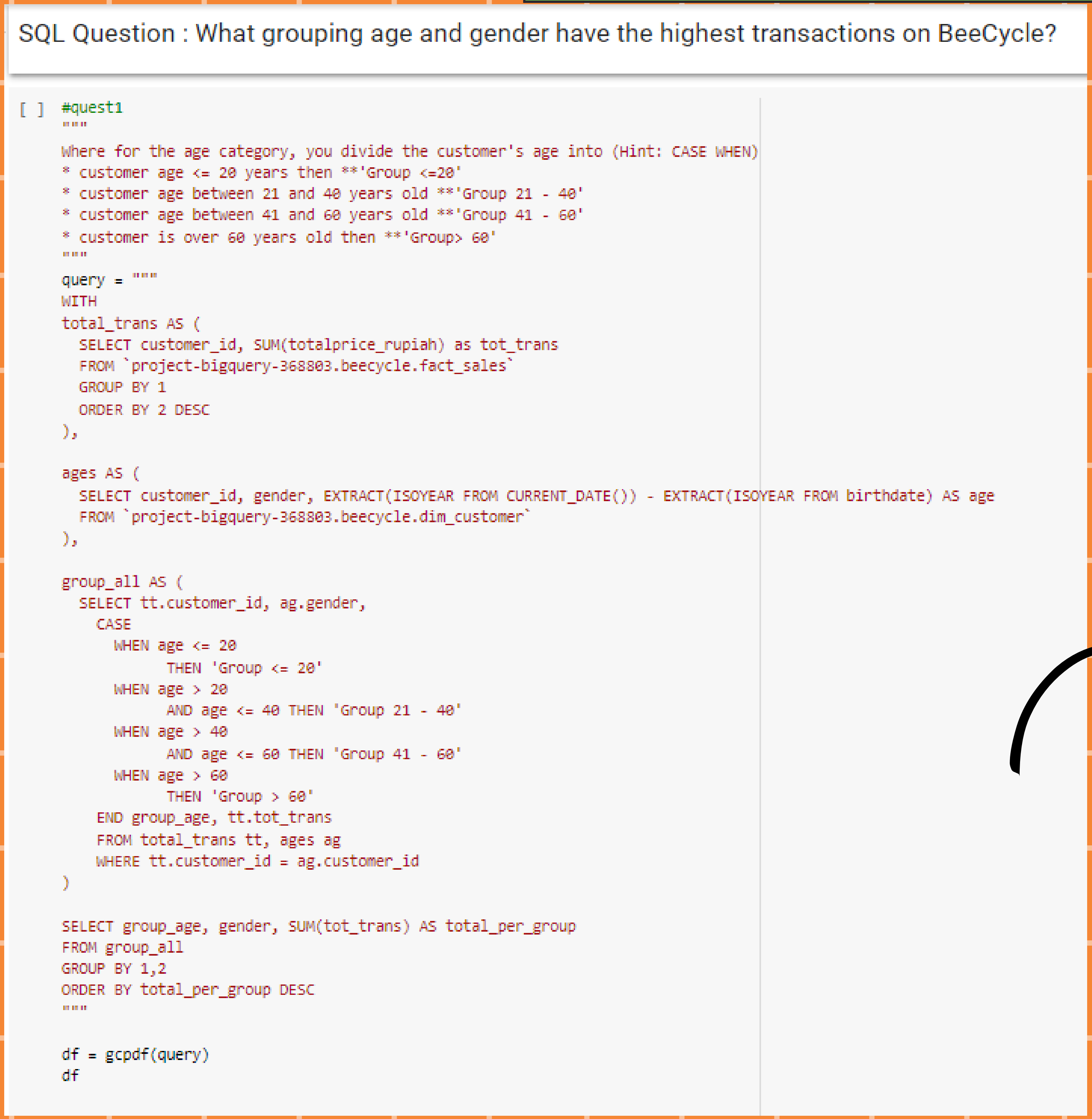


SQL Question : what products are being sold ?

```
[ ] #using select and get value from dim_product and limit 10
    query = """
    select *
    FROM `project-bigquery-368803.beecycle.dim_product`
    LIMIT 10
    """

    df = gcpdf(query)
    df
```

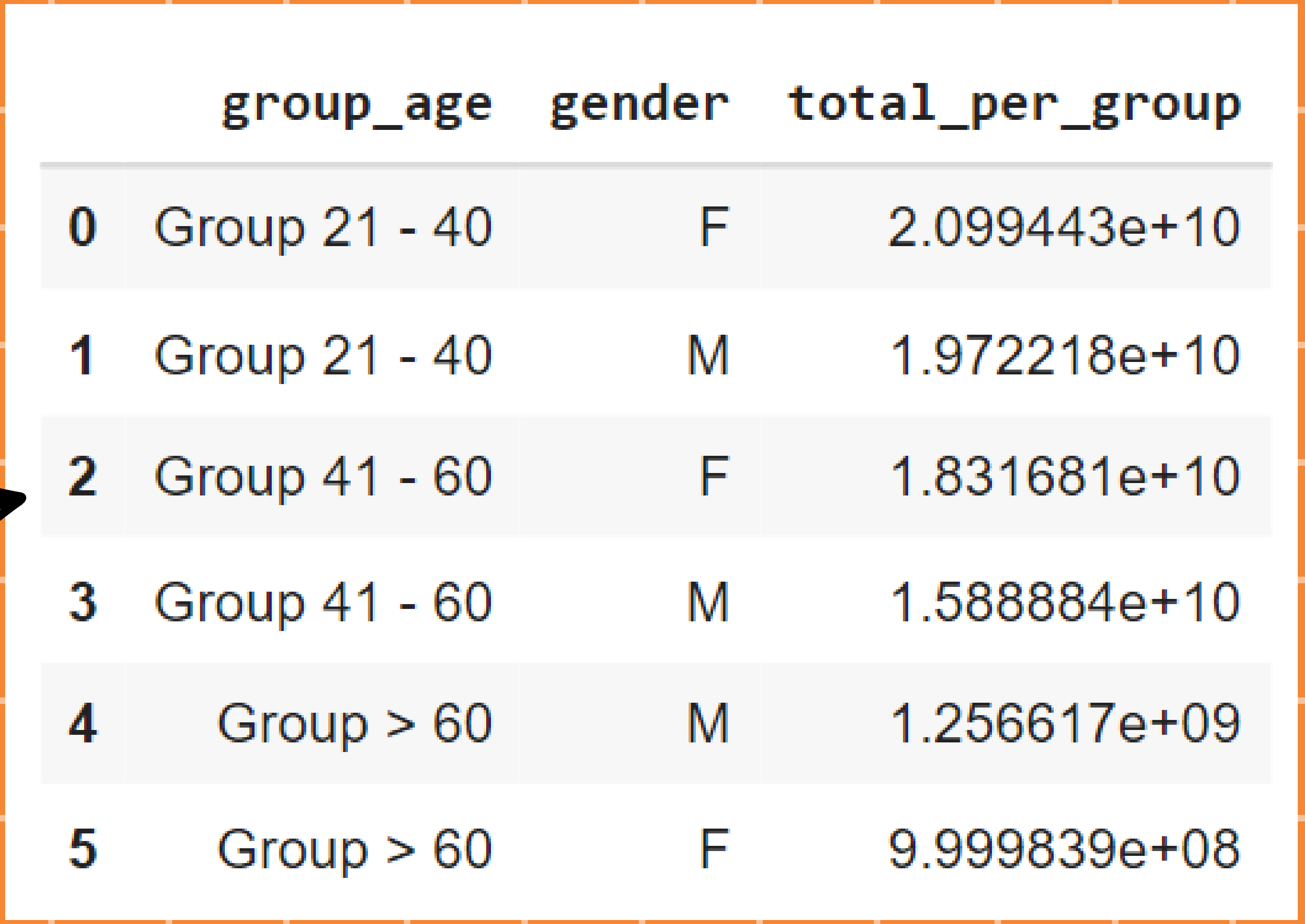
	product_id	product_name	model_name	color	size_range	cost	normal_price	sub_category	category
0	604	Road-750 Black, 44	Road-750	Black	42-46 CM	4811094.4	7559860.0	Road Bikes	Bikes
1	605	Road-750 Black, 48	Road-750	Black	48-52 CM	4811094.4	7559860.0	Road Bikes	Bikes
2	606	Road-750 Black, 52	Road-750	Black	48-52 CM	4811094.4	7559860.0	Road Bikes	Bikes
3	584	Road-750 Black, 58	Road-750	Black	54-58 CM	4811094.4	7559860.0	Road Bikes	Bikes
4	326	Road-650 Red, 44	Road-650	Red	42-46 CM	5784048.2	9787374.8	Road Bikes	Bikes
5	338	Road-650 Black, 44	Road-650	Black	42-46 CM	5784048.2	9787374.8	Road Bikes	Bikes
6	328	Road-650 Red, 48	Road-650	Red	48-52 CM	5784048.2	9787374.8	Road Bikes	Bikes
7	330	Road-650 Red, 52	Road-650	Red	48-52 CM	5784048.2	9787374.8	Road Bikes	Bikes
8	340	Road-650 Black, 48	Road-650	Black	48-52 CM	5784048.2	9787374.8	Road Bikes	Bikes
9	342	Road-650 Black, 52	Road-650	Black	48-52 CM	5784048.2	9787374.8	Road Bikes	Bikes

Try SQL in Colab Using Question



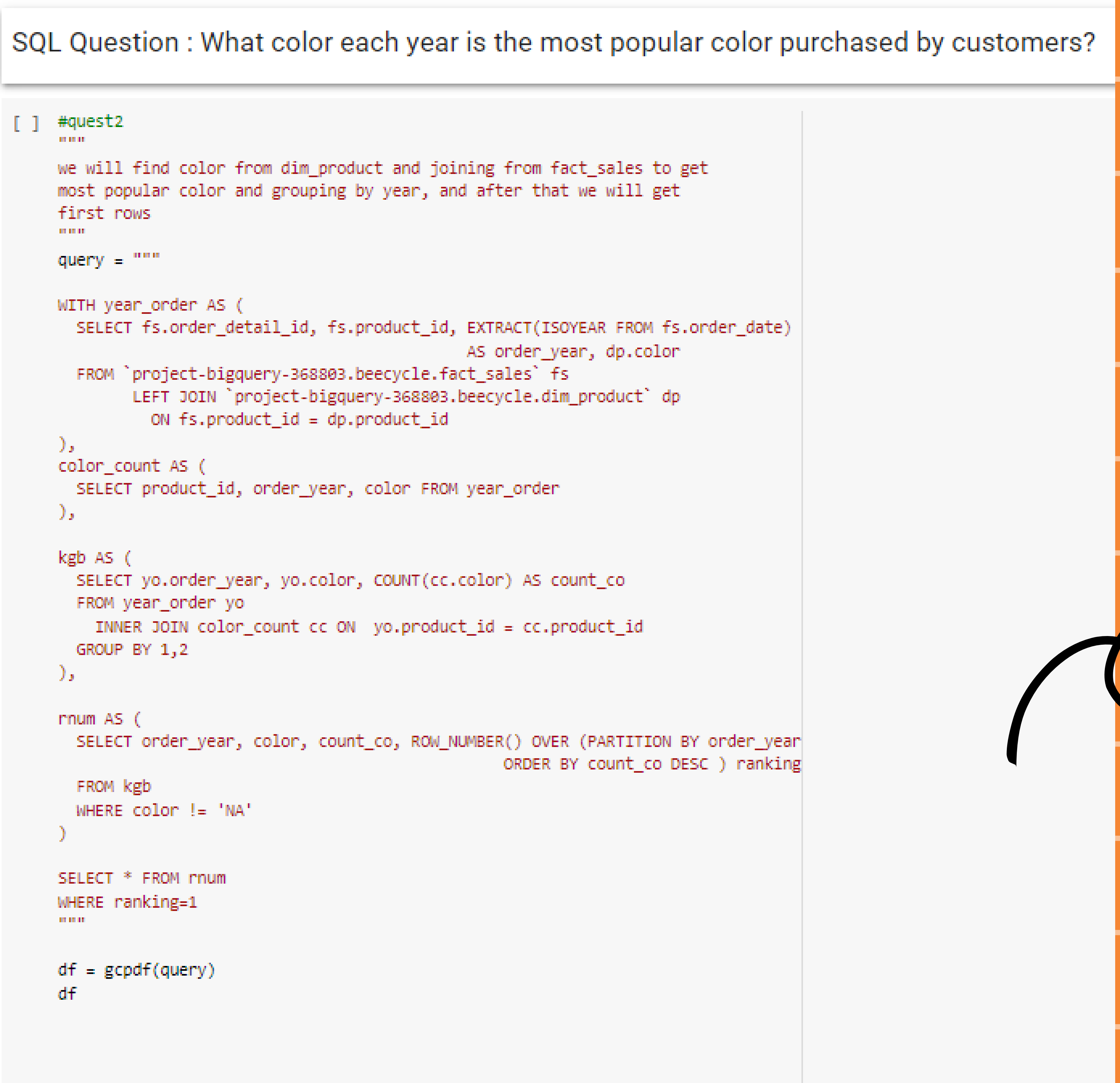
SQL Question : What grouping age and gender have the highest transactions on BeeCycle?

```
[ ] #quest1
    """
    Where for the age category, you divide the customer's age into (Hint: CASE WHEN)
    * customer age <= 20 years then **'Group <=20'
    * customer age between 21 and 40 years old **'Group 21 - 40'
    * customer age between 41 and 60 years old **'Group 41 - 60'
    * customer is over 60 years old then **'Group> 60'
    """
    query = """
    WITH
    total_trans AS (
      SELECT customer_id, SUM(totalprice_rupiah) as tot_trans
      FROM `project-bigquery-368803.beecycle.fact_sales`
      GROUP BY 1
      ORDER BY 2 DESC
    ),

    ages AS (
      SELECT customer_id, gender, EXTRACT(ISOYEAR FROM CURRENT_DATE()) - EXTRACT(ISOYEAR FROM birthdate) AS age
      FROM `project-bigquery-368803.beecycle.dim_customer`
    ),

    group_all AS (
      SELECT tt.customer_id, ag.gender,
        CASE
          WHEN age <= 20
                THEN 'Group <= 20'
          WHEN age > 20
                AND age <= 40 THEN 'Group 21 - 40'
          WHEN age > 40
                AND age <= 60 THEN 'Group 41 - 60'
          WHEN age > 60
                THEN 'Group > 60'
        END group_age, tt.tot_trans
        FROM total_trans tt, ages ag
        WHERE tt.customer_id = ag.customer_id
    )

    SELECT group_age, gender, SUM(tot_trans) AS total_per_group
    FROM group_all
    GROUP BY 1,2
    ORDER BY total_per_group DESC
    """

    df = gcpdf(query)
    df
```



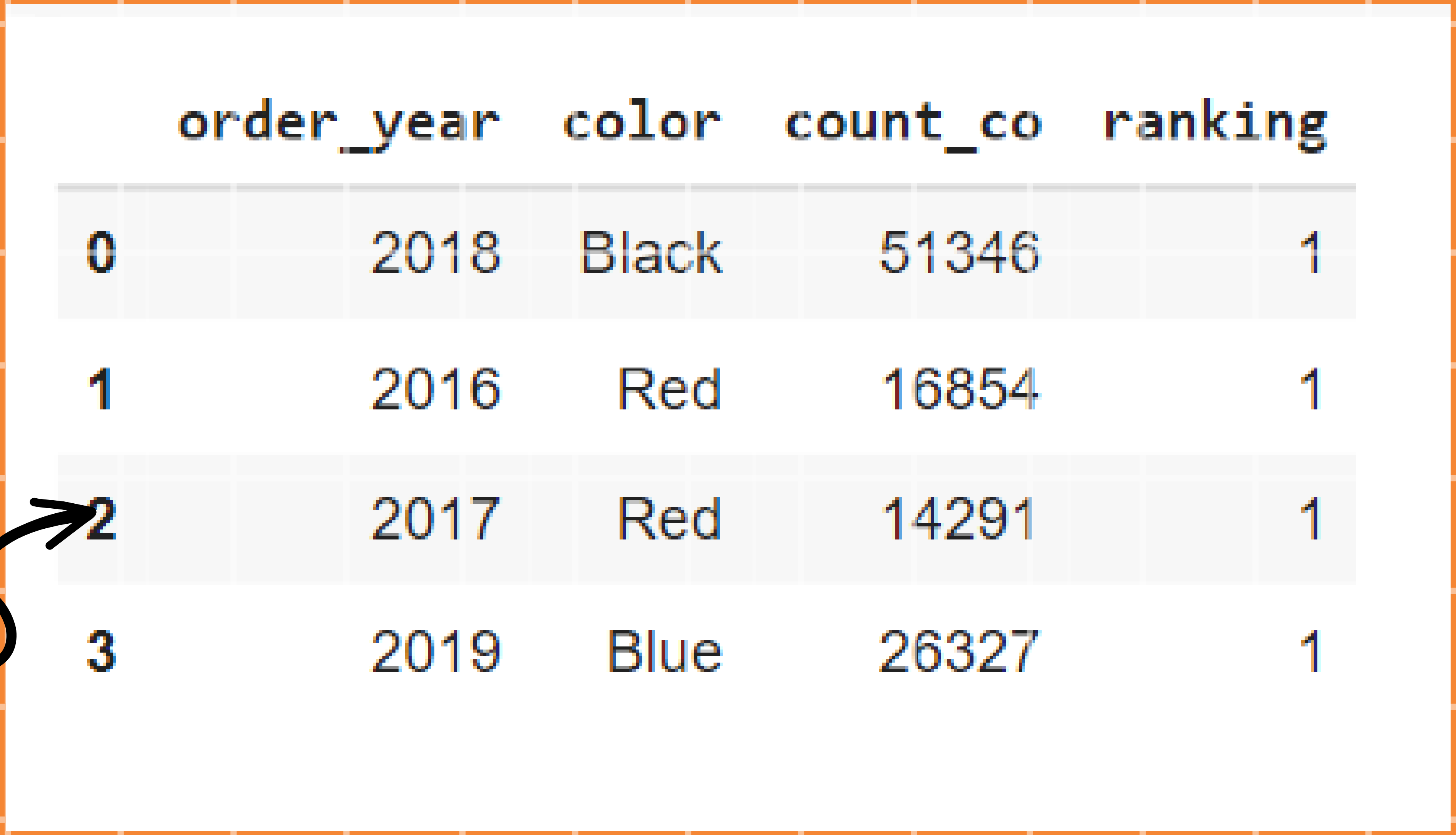
	group_age	gender	total_per_group
0	Group 21 - 40	F	2.099443e+10
1	Group 21 - 40	M	1.972218e+10
2	Group 41 - 60	F	1.831681e+10
3	Group 41 - 60	M	1.588884e+10
4	Group > 60	M	1.256617e+09
5	Group > 60	F	9.999839e+08

Try SQL in Colab Using Question



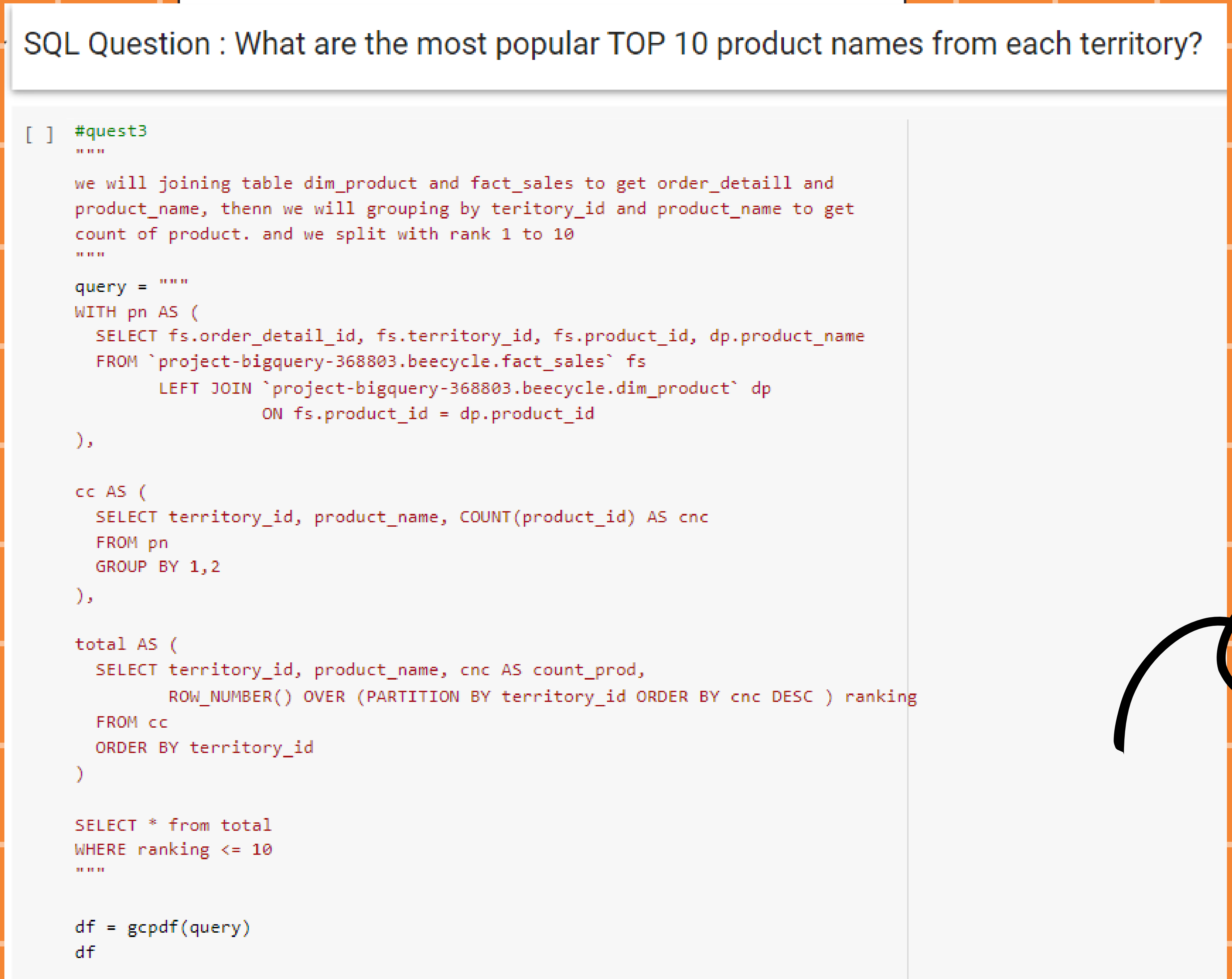
SQL Question : What color each year is the most popular color purchased by customers?

```
[ ] #quest2
    """
    we will find color from dim_product and joining from fact_sales to get
    most popular color and grouping by year, and after that we will get
    first rows
    """
    query = """

    WITH year_order AS (
      SELECT fs.order_detail_id, fs.product_id, EXTRACT(ISOYEAR FROM fs.order_date)
                                                AS order_year, dp.color
      FROM `project-bigquery-368803.beecycle.fact_sales` fs
            LEFT JOIN `project-bigquery-368803.beecycle.dim_product` dp
              ON fs.product_id = dp.product_id
    ),
    color_count AS (
      SELECT product_id, order_year, color FROM year_order
    ),

    kgb AS (
      SELECT yo.order_year, yo.color, COUNT(cc.color) AS count_co
      FROM year_order yo
        INNER JOIN color_count cc ON  yo.product_id = cc.product_id
      GROUP BY 1,2
    ),

    rnum AS (
      SELECT order_year, color, count_co, ROW_NUMBER() OVER (PARTITION BY order_year
                                                     ORDER BY count_co DESC ) ranking
      FROM kgb
      WHERE color != 'NA'
    )

    SELECT * FROM rnum
    WHERE ranking=1
    """

    df = gcpdf(query)
    df
```



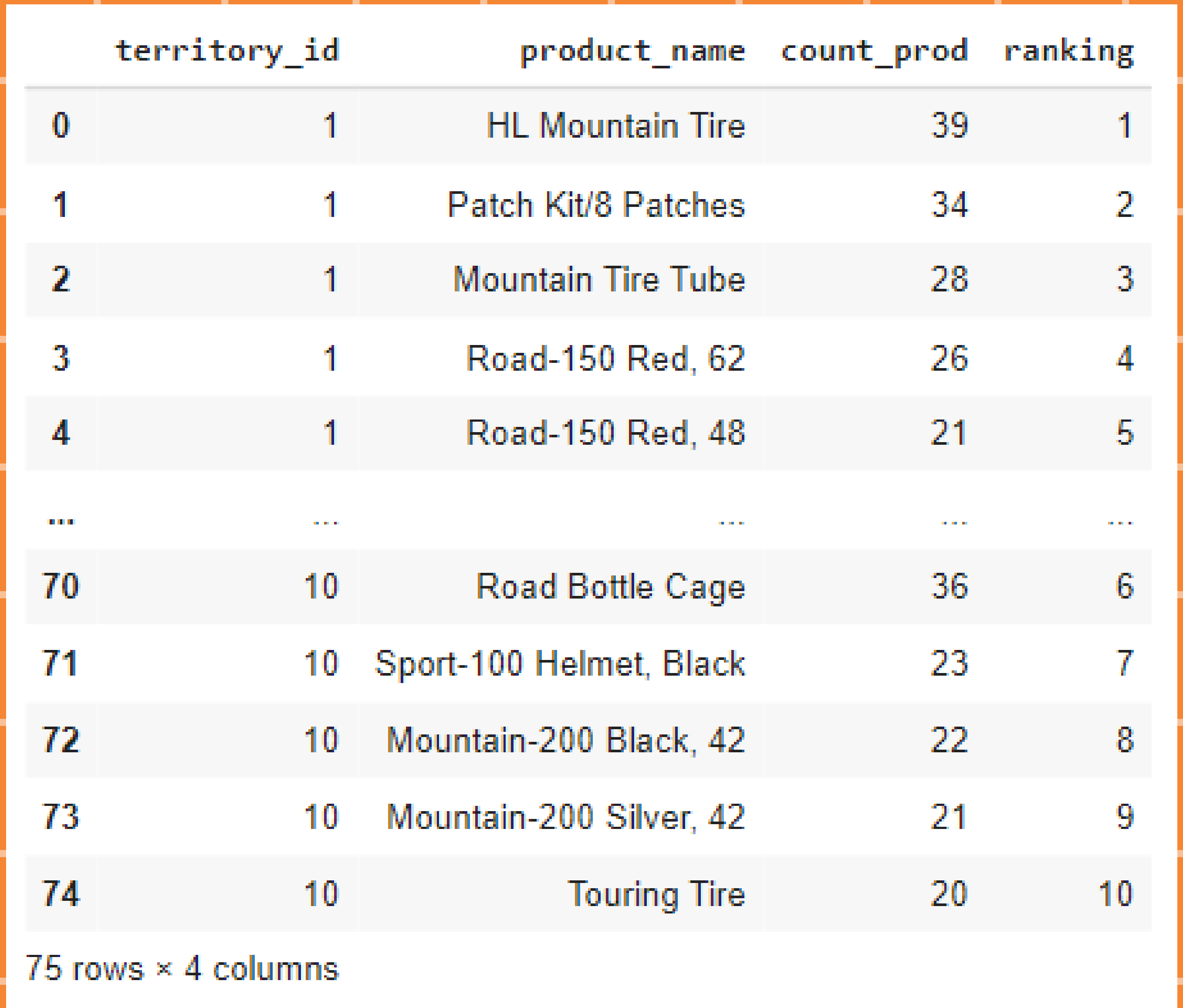
	order_year	color	count_co	ranking
0	2018	Black	51346	1
1	2016	Red	16854	1
2	2017	Red	14291	1
3	2019	Blue	26327	1

Try SQL in Colab Using Question



SQL Question : What are the most popular TOP 10 product names from each territory?

```
[ ] #quest3
    """
    we will joining table dim_product and fact_sales to get order_detaill and
    product_name, thenn we will grouping by teritory_id and product_name to get
    count of product. and we split with rank 1 to 10
    """
    query = """
    WITH pn AS (
      SELECT fs.order_detail_id, fs.territory_id, fs.product_id, dp.product_name
      FROM `project-bigquery-368803.beecycle.fact_sales` fs
            LEFT JOIN `project-bigquery-368803.beecycle.dim_product` dp
                      ON fs.product_id = dp.product_id
    ),

    cc AS (
      SELECT territory_id, product_name, COUNT(product_id) AS cnc
      FROM pn
      GROUP BY 1,2
    ),

    total AS (
      SELECT territory_id, product_name, cnc AS count_prod,
             ROW_NUMBER() OVER (PARTITION BY territory_id ORDER BY cnc DESC ) ranking
      FROM cc
      ORDER BY territory_id
    )

    SELECT * from total
    WHERE ranking <= 10
    """

    df = gcpdf(query)
    df
```



	territory_id	product_name	count_prod	ranking
0	1	HL Mountain Tire	39	1
1	1	Patch Kit/8 Patches	34	2
2	1	Mountain Tire Tube	28	3
3	1	Road-150 Red, 62	26	4
4	1	Road-150 Red, 48	21	5
...	...	...	...	...
70	10	Road Bottle Cage	36	6
71	10	Sport-100 Helmet, Black	23	7
72	10	Mountain-200 Black, 42	22	8
73	10	Mountain-200 Silver, 42	21	9
74	10	Touring Tire	20	10

75 rows × 4 columns

For More Detail

Notebook in Google Colab

bit.ly/3tDywo8

Notebook & Dataset in Github

bit.ly/3ULEFe0

Thank You